These are notes from the Sun Lustre Center of Excellence (LCE) at Oak Ridge National Laboratory, held February 7-8 in Burlington, Massachusetts.

The meeting vision, expected outcomes, agenda, and attendees are listed immediately below, and then notes follow.

Contact the editor of these notes Daniel.Ferber@lustre.org with corrections and additions.

The Meeting's Vision

- Create a far-reaching, strategic vision to bring Lustre to the next level for high-end HPC customers
- Get feedback from the 10 biggest Lustre sites in the world and from strategic partners on how Lustre needs to evolve the next 5-10 years
- Guide Lustre to support hundred Petaflop systems with thousands of storage servers managing an exabyte of data

The Meeting's Expected Outcomes

- Engage the Lustre HPC community in visionary discussions about the future of the technology sites to communicate where they're going and what they need from Lustre
- Establish a community road map for Lustre in HPC in the next 5 years
- Advise the Lustre community on Sun's own plans for Lustre in the next 5 years

Agenda - Thursday February 7th

- Meeting Welcome and Purpose
  8:30am-8:45am, Peter Braam

- Meeting Logistics
  8:45am-9:00am, Dan Ferber

- Introduction and Lustre Key Strategic Topics
  9:00am-11am, All
  Each customer/partner will introduce themselves and talk to one slide, in very briefly reviewing their top 3 topics.
    - *10 minute (timed) for each customer/partner.*
    - *If there are multiple attendees from a specific customer or partner, they should present as one team, and fit into the single time slot.*

- Break
  11am-11:30am

- Lustre Key Strategic Topics (continued)
  11:30am-12:30pm, All

- Lunch (at onsite Sun Cafeteria), 1 hour
  12:30pm-2:00pm, (Each person pays for their own meal)

- Discuss Highest Ranked Topics, in Order of Ranking
  2:00pm-4:30pm, All

- Dinner at Bedford Glenn Hotel* (See note below)
  7pm

Friday February 8th

- HSM at CEA, Special Topic
  8:30am-9:30am, Jacques-Charles Lafoucriere

- Discuss Highest Ranked Topics, in Order of Ranking (continued)
  9:30am-10:30, Discussion

- Break
  10:30am-11am

- Discuss Highest Ranked Topics, in Order of Ranking (continued)
  11am-12:30pm

- Lunch
  12:30pm-2pm, (Each person pays for their own meal)

- Meeting Summary
  2pm-2:30pm, Eric Barton

- Meeting Close and Final Comments
  2:30pm-3pm, Peter Bojanic

Attendees

Bull
Pascale Rosse-Laurent (Pascale.Rosse-Laurent@bull.net)

CEA
Jacques-Charles Lafoucriere (jc.lafoucriere@cea.fr)

Cray
John Carrier (carrier@cray.com)
Jim Harrell (ejh@cray.com)

Data Direct Network
Jeff Denworth - (jdenworth@datadirectnet.com)

DOD
Howard (Flash) Gordon (flash@super.org)

Instrumental
Henry Newman (HPCS) (hsn@instrumental.com)

Lawrence Livermore
Mark Gary (mgary@llnl.gov)
Matt Leininger (leininger4@llnl.gov)

Lustre/Sun Team
Eric Barton (eric.barton@sun.com)
Bryon Neitzel (bryon.neitzel@sun.com)
Peter Braam (peter.braam@sun.com)
Nicole Brown (nicole.brown@sun.com)
Peter Bojanic (peter.bojanic@sun.com)
Kevin Canady (kevin.canady@sun.com)
Stephen Cranage (stephen.cranage@sun.com)
Nikita Danilov (Nikita.Danilov@Sun.COM)
Andreas Dilger (andreas.dilger@sun.com)
Oleg Drokin (oleg.drokin@sun.com)
Dan Ferber (daniel.ferber@sun.com)
Jessica Johnson (jessica.johnson@sun.com)
Karen Jourdenais (karen.jourdenais@sun.com)
Larry McIntosh (Larry.McIntosh@Sun.com)
Alex Tomas (alex.tomas@sun.com)
Ted Wueste (ted.wueste@sun.com)

NASA

Bob Ciotti - (Bob.Ciotti@nasa.gov)

NERSC
Greg Butler (gbutler@nersc.gov)
Bill Kramer (wtkramer@nersc.gov)

NRL
Jim Hoffman (jhofmann@cmf.nrl.navy.mil)

ORNL
Al Geist (gst@ornl.gov)
Shane Cannon (canonrs@ornl.gov)

PNNL
Evan Felix - Evan.Felix@pnl.gov
Rob Farber - rob.farber@pnl.gov

PSC
Brian Johanson - (bjohanso@psc.edu)

Sandia
Steve Monk (smonk@sandia.gov)

Science and Technology Associates
Delores Shaffer (HPCS) - (dshaffer@stassociates.com)

System Fabric Works
Bill Boas (NRL) – (bboas@systemfabricworks.com)

Peter Braam Introductory Comments

Agenda

- Lustre (and Linux HPC software) in Sun
- Deployment successes, requirements & futures
- Product vision

Peter Braam Introductory Comments – Lustre at Sun

CFS acquisition

- Oct 1 the Sun acquisition of CFS closed.
- The theme is continuity
  - > Lustre remains open source under GPL
    - > Today all designs & internals course are on lustre.org
    - > CVS open, architecture discussion now on lustre-devel
  - > Sun continues to work with CFS' partners
    - > No partners lost: DDN, HP, Bull, Cray ...
    - > No special versions of Lustre for anyone
  - > No customers were lost
  - > No employees lost

Lustre Team

- Mostly similarly structured
  - > Support for large sites follows CFS model
- Changes
  - > Braam: Chief Architect – customer reqs / product vision
  - > Bojanic: Manages team – support & project planning
  - > Barton: CTO - implementation lead
  - > Marketing & sales moved to other groups

New effort – Sun HPC Software

- In Bojanic's group
- An element of a bigger HPC effort in Sun

What is in this solution?

- Existing Linux Distributions
  - > Not a new Linux distribution
- HPC packages from
  - > Sun: Grid Engine, Lustre, QFS, Sun Studio, Clustertools
  - > the community: numerous
  - > ISV's: TotalView, LSF …
- A test suite to validate installations
  - > Handle hardware & software substitutions
- Aligned with Solaris HPC bundles

Peter Braam Introductory Comments – Lustre at Sun

What will Sun offer?

- Sun will certify the software on Sun hardware
  - > Turnkey HPC solutions
- Sun will support the solution
  - > Including Linux & IB. Sometimes with partners
  - > Training & documentation
- The certification suite will be open source
- The effort will be mostly visible to the community

Lustre deployments today

- Top 500 Lustre share
  - > 7 of top 10, 50% of top 30, 30% of top 100
  - > Multi cluster Lustre: LLNL, ORNL, Sandia
  - > Wide area initiatives: DoD, TeraGrid
- Partners
  - > Bull, Cray, HP, Dell, Hitachi, DDN, Sicortex, Terascala,…
- Commercial deployments
  - > Growing – Oil & Gas, EDA, Manufacturing, Media, ISP's
- Business Strategy – Lustre is for extreme storage

Peter Braam Introductory Comments – Lustre Deployments

How have things gone?

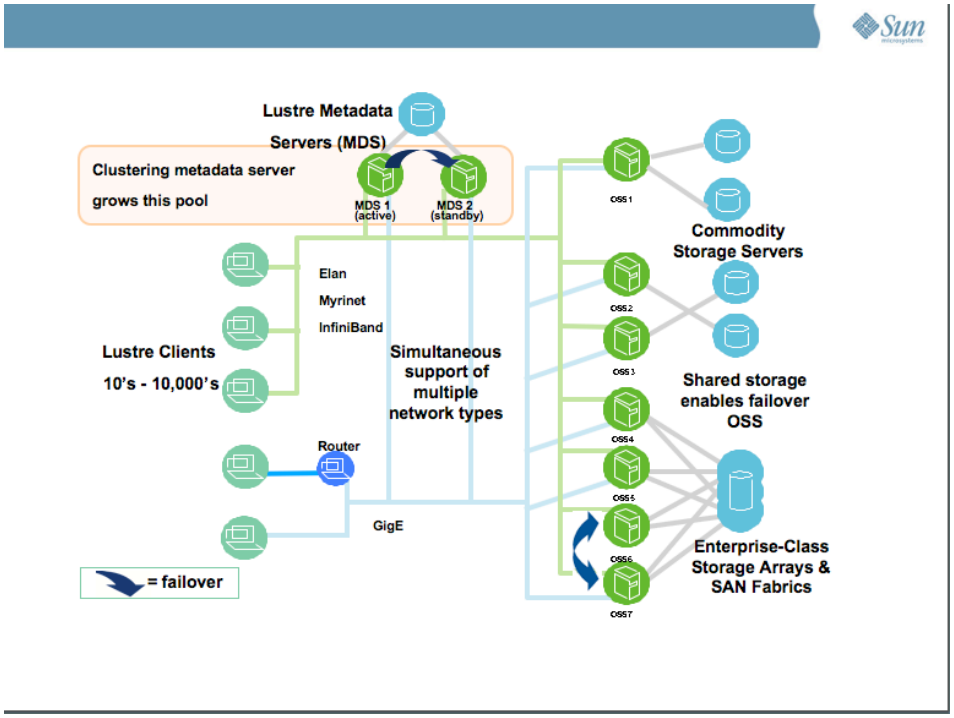
Issue	Result
The most scalable HPC FS	Good – 5 years in a row now, 7 of the top 10
Offering high product quality	Improving, but far from a Skype or OS X like experience
Broad adoption	Not yet, not on track for it

Vision

Facet	Activity	Difficulty	Priority	Timeframe
Product Quality	Major work is needed, except on networking	High	High	2008
Performance fixes	Systematic benchmarking & tuning	Low	Medium	2009
More HPC Scalability	Clustered MDS, Flash cache, WB cache, *Request Scheduling*, Resource management, *ZFS*	Medium	Medium	2009 - 2012
Wide area features	*Security*, WAN performance, proxies, replicas	Medium	Medium	2009 - 2012
Broad adoption	Combined pNFS / Lustre exports	High	Low	2009 - 2012

Note: These are visions, not commitments

Peter Braam Introductory Comments – Vision (not commitments)



Lustre will get ZFS DMU backend

- Current server implementation
  - Servers are in ***patched*** Linux kernel
  - We require ***modified*** **ext4 (ldiskfs)**
- Customers require
  - Portability
    - No kernel patches
    - Platform independent API
  - Scalability
  - Hardening
  - To get this all into ext4
    - Estimated @24 FTE years
- We explored
  - User space servers on FS
    - MDS not possible
  - Harden & scale ext4
    - Possible – effort too high
    - Community going too slow
- We choose an alternative
  - Use ZFS DMU
  - Pro: can be in user space
  - Pro: scalable, hardened, portable
  - Pro: much less work (est 3x less)
  - Con: performance work

Peter Braam Introductory Comments – Vision (not commitments)

Lustre & data integrity

- ZFS DMU has storage integrity
  - > Lustre will use it
- Lustre adds network integrity
  - > Compare data before & after network DMA
- When the feature was added
  - > It discovered a few data corrupting Elan3 cards!
  - > Nobody new about this

Network Request Scheduler

- Fileservers today process a request queue as fifo
- The NRS will re-order requests and…
  - > Allow clients to make fair progress
  - > Re-order I/O to make it sequential in the diskfs
  - > Pre-fetch metadata to avoid blocking
- Use in conjunction with OST & MDT write caches
- The 2nd gen NRS will add server coordination

Peter Braam Introductory Comments – Vision (not commitments)

Flash cache

- Exploit storage hardware revolution
  - > Very high bandwidth available from flash
  - > Add Flash Cache OSTs– capacity ~ RAM of cluster
  - > Cost: small fraction of cost of RAM of cluster
- Allow fast I/O from compute node memory to flash
- Then drain flash to disk storage - ~ 5x slower
  - > E.g. cluster finishes I/O in 10 mins, on disk in 50 mins
  - > Need 5x fewer disks
- Lustre manages a coherent view of the file system

Metadata WBC

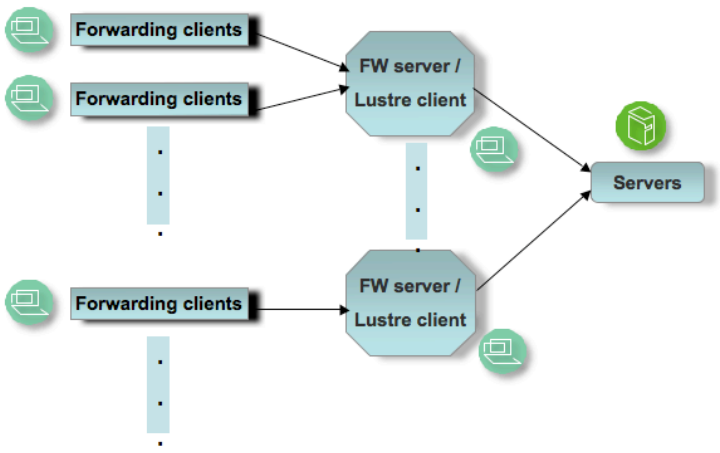
- Goal & problem:
  - > Disk file systems make updates in memory
  - > Network FS's do not - metadata ops require RPCs
  - > The Lustre WBC should only require synchronous RPCs for cache misses
- Key elements of the design
  - > Clients can determine file identifiers for new files
  - > A change log is maintained on the client
  - > Parallel reintegration of log to clustered MD servers
  - > Sub-tree locks – enlarge lock granularity

Peter Braam Introductory Comments – Vision (not commitments)

Uses of the WBC

- HPC
  - > I/O forwarding makes Lustre clients I/O call servers
  - > These servers can run on WBC clients
- Exa-scale clusters
  - > WBC enables last minute resource allocation
- WAN Lustre
  - > Eliminate latency from wide area use for updates
- HPCS
  - > Dramatically increase small file performance

Lustre with I/O forwarding

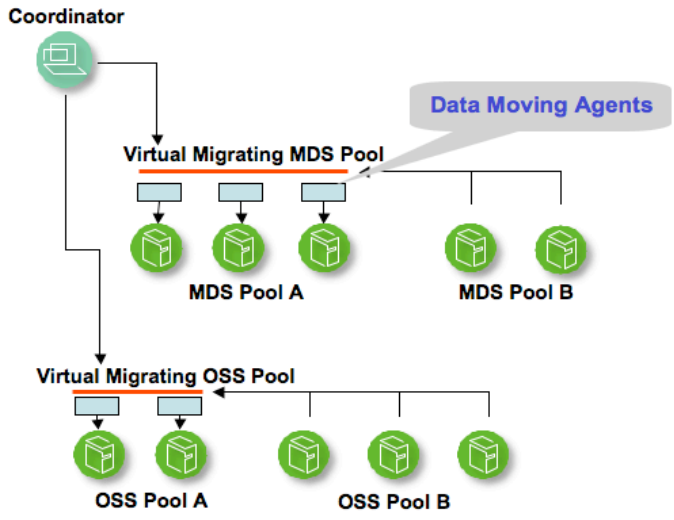


Peter Braam Introductory Comments – Vision (not commitments)

Migration – many uses

- Between ext3 / ZFS servers
- For space rebalancing
- To empty servers and replace them
- In conjunction with HSM
- To manage caches

Migration

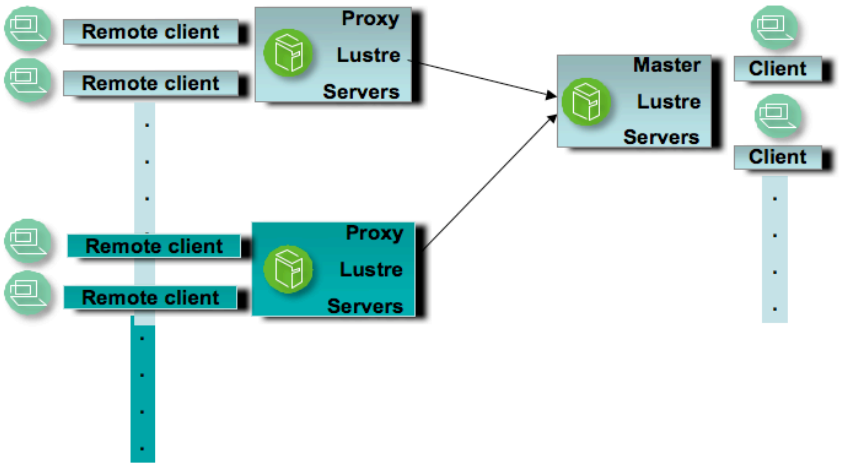


Peter Braam Introductory Comments – Vision (not commitments)

Caches / proxies

- Many variants
  - > HSM – Lustre cluster is proxy cache for 3rd tier storage
  - > Collaborative read cache
    - > Bit-torrent style reading or
    - > When concurrency increases use other OSS's as proxies
  - > Wide area cache – repeated reads come from cache
- Technical elements
  - > Migrate data between storage pools
  - > Re-validate cached data with versions
  - > Hierarchical management of consistency

Proxy clusters



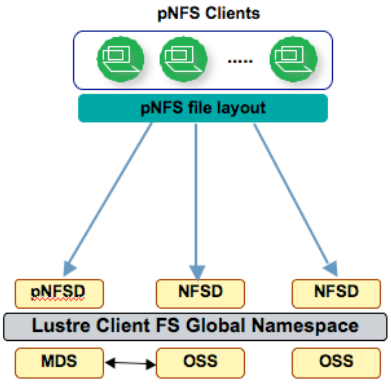
Local performance after the first read

Peter Braam Introductory Comments – Vision (not commitments)

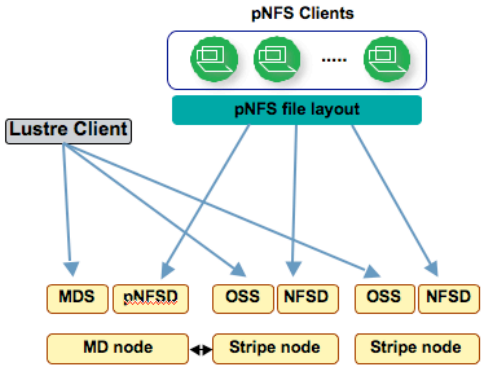
Broad adoption vision

- pNFS integration
- Soon – pNFS exports from Lustre on Linux
- Longer term
  - > Let Lustre servers offer pNFS & Lustre protocol
    - > Requires an interesting Lustre storage layer
  - > Make LNET an RDMA transport for NFS
  - > Offer proven Lustre features to NFS standards efforts

Layered & direct pNFS



pNFS layered on Lustre Clients



pNFS and Lustre servers on Lustre / DMU storage system

Major Themes of Importance
Presented by Each Customer or Partner

TOP 4 for Lustre Five years Priorities

SECURITY

1) Large Cluster Lustre File system
   - Management of a large configuration :
     - Initialization large Number OST with template
     - Correlation Storage configuration / File system Configuration
   - Support of multiple concurrent failure

2) Data Integrity
   - Decrease Recovery time before a restart on a valid FS
   - Error Diagnostic Enhancement.

3)Wide area Lustre
   - Lustre capability to share data over a distributed community

4) HSM
   - Optimize Storage Space attach to the compute cluster
   - Integration to the batch scheduling

- Discussion Comments
  - Large config management, hard to debug failures, especially double failures
  - Decrease recovery time before a restart
  - WAN – share data over distributed community
  - Security applies to all these topics

Customer and Partner Strategic Priorities

CEA top topics

0) *Allow Lustre to connect transparently to external storage*

- It is a short term work so we do not need to prioritized it for the long term (this point is just a reminder)

1) Guide Lustre to support hundred Petaflop systems with thousands of storage servers managing an exabyte of data

- Next CEA clusters will be Petaflops systems around 2010-2011and they will continue to grow (x10 every 4 years)

2) Multi protocol version/release support

- We will put Lustre at the center of CEA computing centers and all clients will not evolve synchronously

3) Storage technology and Lustre

- We want to be able to choose between different storage technologies on which Lustre is optimized

4) Advise the Lustre community on Sun's own plans for Lustre in the next 5 years

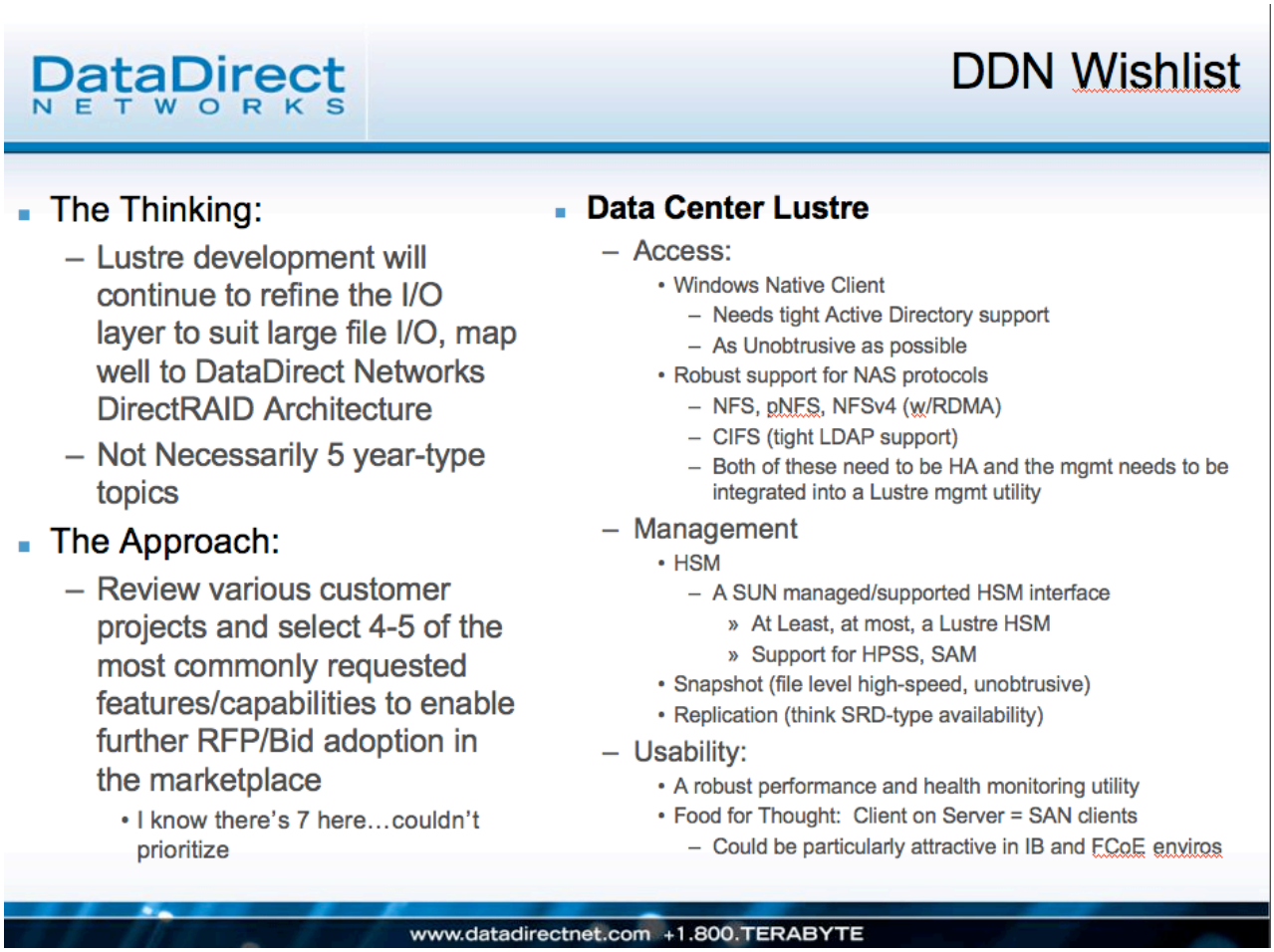
- We need to know the Lustre "natural" evolution

- Discussion Comments
  - 10x growth every 4 yrs.
  - Multi-cluster compatibility, interoperability is key, don't want to use pNFS
  - Open disk choice
  - Need to know the Lustre roadmap – not necessarily Sun's corporate roadmap

Lustre Discussion Topics

- Production Quality Lustre
  - Release Schedule – stable and structured
  - Compatibility – backward and forward
  - Quality – metrics and tracking
  - Better Testing – too many errors found at customer sites
  - Problem Analysis – diagnose on first failure
- Performance Improvements and Tracking
  - Test Suites – specify known set of tests with reproducible results
  - Processing Overhead – reduce the "cost" of Lustre on clients and servers
- Scaling
  - Goal – scale to 50K nodes
  - Measurability – define methods for measuring scalability (e.g., HPCS Scenarios)
  - Limits – know from design (e.g., transactions/sec on a given HW/SW combination)
- Future Directions
  - Enhancements – customers expect backup, HSM, and complementary features, which should leverage and be compatible with existing solutions
  - Sun's Goals – is Lustre a cluster file system, a scalable file system, or a full featured file system for any purpose? All of the above?

- Discussion Comments
  - Production quality, problem analysis is important
  - Be able to track performance profile over time
  - Need to know transaction rates based on hw/sw
  - Where is Lustre going? Focused still on HPC?



- Discussion Comments
  - Windows client, Active Directory enabled, works in the data center
  - Need HSM solution – CEA solution needs to be Sun supported
  - Replication (SRD-type availability)
  - Monitoring
  - Client on Server = SAN clients
  - Sun's Samba doesn't support AD. Bojanic says we still recommend using Linux based Samba. OSR will be doing the native Windows client, will talk to them about Active Dir capability.

DoD

- Performance for large and small files
- Reliability
- Recovery of MDS after a crash.

- Discussion Comments
  - Interested in scalability and data integrity
  - How long to reload
  - Reliability – T10 DIF –equivalent – 10x17 read error bit – 10x28 undetectable or mis-corrected errors
  - Don't want to spend time on repairs, don't bring system down. T10 DIF sets the standard.
  - Is ZFS is as good or better than T10 DIF? Some thought so. Sun will review with Henry. Sun not interested in special disks with error correction. See a push toward commodity storage. Need agreement on requirements and specifications.
  - Don't care how the requirement is satisfied, T10 DIF or other is fine.
  - Do we need to point to the error? Yes. ZFS only points to error on the server. If client reports error without pointing to error, this is not good. Diagnosis vs. checking could be an issue.

HPCS Requirements

- **1 Trillion files in a single file system**
  - **This has implications for the file system and Linux**
- **32,000 file creates per second**
  - **This has meta data design impacts along with hardware architecture impacts**
- **10,000 metadata operations per second**
  - **Things that call stat**
- **Streaming I/O at 30 GB/sec full duplex**
  - **Needed for real-time capture systems**
- **Support for 30,000 nodes**
  - **All writing or reading at the same time**

- Discussion Comments
  - 1 Trillion files, 32K file creates/sec, 10K md ops/sec, lots of ls –l, 30GB/sec from one client, 30K nodes
  - Roadmap visibility is important, quality and schedule are important
  - Must be POSIX compliant, maybe the Posix std should be changed
  - No Benchmarking Tricks. DARPA wants linear, predictable scalability
  - Quality and schedule are harder than meeting the scalability targets. There is no change with the Lustre relationship with the Linux community.

LLNL Priorities – and what we think they mean…

- **Ultra-large configurations and how to support them**
  - Scaling in all dimensions
  - MDS Performance - CMD implementation, SSD-based MDS
  - Multi/many-core parallelization
  - *High performance* ZFS implementation
  - Free space management
- **Data Integrity at the petascale to the exascale**
  - ZFS (lack of fsck, checksums…)
  - Lustre RAID
- **Wide Area Lustre (QoS; guaranteed bandwidth; latency)**
  - *For QoS* – this hurts us today

- Discussion Comments
  - Lots of FPP applications. Need better MDS performance
  - Cores aren't getting faster. Need more parallelism
  - ZFS is high priority, but can't go backwards on performance
  - Admin is important. Free space management.
  - RAS is key. We don't want the 3am calls!
  - Lustre RAID would be great. Don't want to rely on failover
  - Need QoS - one viz guy can impact everyone. Will get one straggler node
  - Even if knobs are available, how to use the controls? Make it automatic
  - L-RAID 2 years ago would have been good. Re-evaluating in light of the new SAS shared storage. Low cost JBODS w/ SAS interconnects which is shared failover capable storage. Not a simple feature to get right, or make it perform. Will need read-modify-write scenarios in every job.
  - Lack of ZFS fsck can be a big problem. This is in the roadmap, but would not needed very often. Would be online.
  - QoS – how to solve fairness issues (political when dealing with multiple cluster users)

NASA Ames

1) kernel dependencies. Need to address the strict kernel/lustre/openIB dependencies. This can be quite problematic when the kernel release we can run is defined by the latest lustre release. This is bad enough for a single system - then consider a large facility with 1000's of systems. This includes client/server OS rev limitations especially when we will want to export a filesystem to compute, archive, vis and WAN. Managing upgrades is also a big problem - have to support staged upgrades (can't have the whole facility down at any one time).

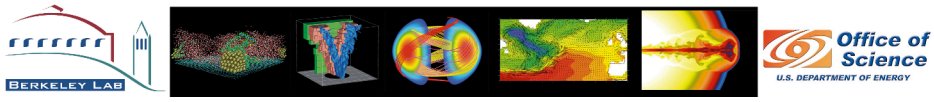
2) large SSI performance. Our largest machine is a fat node cluster. 20 512cpu + 2 1024 + 1 2048 - where the 512/1024/2048 are single system image SGI systems. We have an interest in this going forward, and wan tighter integration with our cluster/vis systems.

3) single thread performance. Maximizing performance of Rank-0 I/O (or few rank I/O) can make porting issues less problematic.

- Discussion Comments
  - 3 domains (clusters) all share data. Each domain has (vis, compute, archive)
  - Does mixing file system and archive system reduce the reliability?
  - Reliability is key. Rely on fs always being there.
  - Must support staged upgrades.

NERSC's Top 5 Priorities

- Support for future High Performance Computing (HPC) systems
  - Support for >> 100k+ clients
  - Support for the major HPC vendor's present and future systems
  - Support for HPC vendor parallel I/O stacks
  - Support fine-grained concurrent access from all clients simultaneously using less than 1KB per client
- Scalability
  - Meta-data performance, Streaming I/O for single stream and multi-threaded, multi-node I/O, IOPS, Storage capacity, Number of clients
- Information Lifecycle Management and HSM Functionality
  - Data layout
  - Online data migration (file, object, and OST) automation, load balancing, restriping
  - Storage pools
  - Data placement policies
  - HPSS integration, for managing data on tape, including HSM
- Backup/Restore capability
  - Fast meta-data scan
  - Snapshots
  - Scalable backup and restore
  - HSM aware integration
- System and Filesystem Administration
  - Fault and performance monitoring
  - Fast parallel filesystem integrity checking and recovery tools
  - Online configuration management and configuration changes to live file systems
  - Programmable interfaces to administration tasks
  - Automated failover and recovery
  - Client authentication and security



- Discussion Comments
  - 100K clients, storage pools, good performance on non-block aligned IO

Lustre Increment Components for JCTD

- Add "Global" capability
  - Namespace, Consistency Semantics
- "Caching" and Data Migration (Replication)
- Windows Native Client for XP and Vista for Lustre networking
- WAN distributed performance
  - Continued OFED compatibility
- Road to PKI/CAC using Kerberos update
- Controlled Interface compatibility
  - Future: Cross-domain compatibility
- Advanced Search and Oracle Interfaces
- Early Transition Observations:
  - A Lustre foundation optimally supports WebDAV 2.0 P2P and Ajax apps which are at the core of RT10/RTRG and MDA
  - Believe we can show that it supports light weight clients as is called for in the DODIIS document.
- Future growth area:
  - For GIG-E and lower users OSS performance can be gained via virtualization technologies and through 4th generation IB hardware and ZEN/VMWare/Hypervisor ("Large Data laptop" technologies)

- *Proposal:* Develop a *"Wide-area Center of Excellence"*
  - *Goal:* Address the widening gap between application and sensor I/O demands and wide-area transport to enable sharing of terabyte flows … from exabytes of data world-wide among federated, distributed "Large Data" portals
    - *e.g., PetaScale HPC, LHC, VLBI, LSST Telescope, Persistent Imagery, …*
  - *Participants:* NRL, NSA/LTS, MIT/LL, UIC National Center for Data Mining, USC/ISI
  - *Testbed:* ATDnet/BoSSnet, TeraFlow/NLR, DREN, …
  - *POPs:* Boston, D.C., Chicago, Los Angeles, Ontario, Amsterdam, Geneva, Tokyo, …
  - *Performance:* 10G, 40G, 100G flows … scaling over time to Terabit flows … UDT OpenSource transport, etc.
  - *Instrumentation:* "Gargoyle" Argus 3.0+ sensor technology, with assist from *FPGA & ManyCore* hardware elements for 10G/40G/100G flow performance E2E

- Discussion Comments
  - Global features, caching and data migration
  - 10Gbs – 100s TB / site
  - Infiniband over WAN – uses Yotta Yotta
  - Windows client (XP, Vista)
  - WAN performance – using OFED
  - CAC – common access card – Kerberos is a path to PKI
  - Data in sync – hi-res vs lo-res based on user's nationality

Top 3 issues for Future Lustre versions

- **1. Evolve Lustre towards a more community driven development model:**

  Note: Sun has recently opened up access to Lustre CVS repository. Beyond this we could envision bringing in competent developers from the labs and vested vendor partners. A more inclusive governance model would be the ultimate step.

- **2. Support for future High Performance Computing (HPC) systems:**
  - **Affordable cost model (DOE site license instead of per client)**
  - **Fully functional clustered metadata server system**
  - **Support for 100k+ clients with 1000+ OSSs**
  - **Support for the major HPC vendor's present and future systems**
  - **Support for HPC vendor parallel I/O stacks**
  - **Support fine-grained concurrent access from**
  - **all clients simultaneously using less than 1KB per client**
  - **Support DARPA metric (30,000 file creates per second)**

Note: To reach these levels would likely require some significant amount of re-architecting (such as proxy servers, cooperative caches, locking enhancements, etc).

- **3. Improved support for multi-clustered environments:**
  - **Heterogeneous networks, routing, software, hardware, and architectures**
  - **Client authentication and security**
  - **Fine-grained QoS support all the way down to the application level**
  -

  In addition, Lustre must accommodate increasing amount of version skew between clients, routers, (proxies), and servers.

- Discussion Comments
  - Community driven development model – keep Lustre open
  - Congestion management – need LNET improvements
  - Need better security model – can't always 'trust your client'

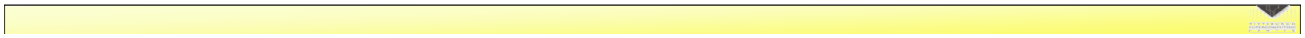
Rob Farber Top 5

1. Production high-performance, high reliability windows and other OS support (either native client or gateway)
2. Get rid of Meta-data bottlenecks (scaling/robustness)
3. Remove security holes
4. WAN support (adapt/configure for high-latency and bandwidth long-haul networks, security, multi-institution ACLs, etc.)
5. Production ultra-wide striping support (with ability to transparently remove or migrate out slow or failing devices)

- Discussion Comments
  - Have hi-speed instrumentation machines – need native Windows clients or high performance gateway.
  - Have MD bottlenecks – scaling/robustness issues – 25MB/s per Gigaflop
  - Remove security holes – can one group compromise another group's data
  - WAN support – adapt for hi-latency, need production capable hi-bandwidth
  - Multi-institution ACLs
  - Ultra-wide striping support
  - 3K OSTs 126GBps Read perf. Only as fast as the slowest device; migrate out slow or failing devices
  - Will have 25K disk drives soon.
  - HSM – need secondary copies of data. DoE lab doesn't control what goes in the data center. Need HSM to work with lots of devices.

PSC – LCE Summit

- Wide area lustre
  - PSC is funded by the NSF Teragrid project to explore the feasibility of using Lustre as a Teragrid-wide, global filesystem
  - The Teragrid is currently using IBM GPFS at selected sites, widespread adoption is limited by licensing issues with IBM
  - Performance of lustre over WAN has been tested PSC to NCSA, ORNL, and NCRA
  - Administration issues need to be addressed: Authentication, User account management
- Hierarchal storage management
  - Closer integration between lustre and arbitrary HSMs
  - PSC requires an overt action from users to move their data to the HSM, we'd like it to be non-transparent
- Lustre and parallel IO middleware
  - PSC developed IO middleware Zest, a transparent intercept library to the client
  - Complement to lustre, not a full filesystem, tuned for one purpose
  - It sits between the production system and Lustre, giving up to 93% of peak aggregate spindle speed
  - Zest breaks up the incoming data into chunks sized for efficiently writing to disk. JBOD arrays are managed by server software with one process per disk.
  - Extensive error checking, recoverability, and monitoring are built into the package.



- Discussion Topics
  - WAN – PSC funded by NSF Teragrid. Authentication and User account mgmt
  - Don’t want to be nailed down to a specific HSM
  - Zest – integrate this with Lustre. Write-only FS.

SNL's top three priorities:

- storage technology and Lustre (new disk storage; flash; etc)
- ultra-large configurations and how to support them (e.g. 5000 OSTs)
- data Integrity at the petascale to the exascale

Bonus item:
- Lustre on other platform (Windows, Macintosh, etc.)

- Discussion Comments
  - Would like to see Lustre appliance.
  - Better failover, human readable error messages
  - Data integrity, ZFS looks promising, 300TB fsck took 15mins
  - L-RAID is important, part of appliance story. Might take perf hit, but that's ok to a degree
  - Would like native Windows client

Consolidated and Ranked Topics
Rankings are only for the purposes of ordering the discussion. We all recognized that all topics were highly important.

System and Filesystem Administration (includes Usability)
Improved support for multi-clustered environments (including QoS)
Data Integrity
Evolve Lustre towards a more community driven development model
Support for Very/Ultra-Large Large Clusters and WAN
Production Quality Lustre
Multi protocol version/release support
Security
Information Lifecycle Management and HSM Functionality
Backup/Restore capability
Performance Improvements
Usability of Lustre
Get rid of Meta-data bottlenecks
HPCS Requirements
Production high-performance, high reliability windows/other OS support
Lustre and parallel IO middleware
External Storage Flexibility, and New Storage Support
Wide Area Center of Excellence
Production ultra-wide striping support

The group then went back and talked in more detail on the above items, in this order. Those notes follow.

System and Filesystem Administration (includes Usability)

- Discussion Comments
- Free Space Management
    - Need to drain an OST
    - Rebalancing of existing files
    - PNNL wants to fill one OST at a time.
    - PSC questions how important this is relative to the other topics
    - Sandia says running out of space kills jobs at inopportune time
    - Andreas says Lustre could provide the low level verb, but users should script the daemons to manage each sites policy
    - Need to drain OSTs, for a variety of reasons
    - Instrumental says Harriet Coverston is already doing this
- Need better diagnostics
- NERSC monitors network traffic to determine if an OST has gone silent
- Bojanic – how many people use LMT? ORNL does. Maybe it would be used more if it was in a Lustre rpm. LMT is too specific to LLNL and Chaos – Sandia concurs. LMT maintainers at LLNL were lost in budget cuts.
- CEA – SNMP support should be extended. /proc is moving to .lustre. Need a single SNMP daemon to collect all information. Andreas – not many people ask for SNMP. CEA – Lustre should do instrumented data aggregation. Big problem is straggler. Need an easy way to find this out. Need tools to find slow hw, processes, etc.
- Could use grad students or luster.org community to build off hooks in Lustre.
- NERSC uses Cacti. Others have mentioned Ganglia, CollectL
- Multi-cluster management is also an important topic
- NERSC doesn't always like Kerberos – too hard to implement. Want a lighter weight solution.
- NERSC – might need multiple plug-in modules and policies for different cases within one site
- NERSC doesn't use firewalls. Uses intrusion detection. Systems don't trust each other.

System and Filesystem Administration (includes Usability) - continued

Failover recovery discussion comments

- GPFS has bullet proof recovery. Needed at OSS, OST level, transparent to users. Controllers, switches, line cards, servers are all types of failures. 12K nodes – 200 days w/o user visible failure. 15 second failover. A function of how tight heartbeat is set. Only lost data once in 2 years.
- Data integrity was on many lists in the morning.
- In 1.6.1 checksum is on, 1.6.2-4 is off. 1.6.5 will have it on again. Will have checksum from write request to disk.
- There are tradeoffs to check summing.
- PNNL runs manual failover. CEA runs auto failover. Works most of the time. LLNL doesn't even run manual failover. Almost no one runs failover. Those that do usually run manual.
- Double mount protection is critical to running auto Failover.
- Parallel backup, high speed metadata scan with file sizes – no one else brought this up? Want to restore the directory and let HSM restage as needed.
- HSM restore can't use physical information from the failed file system. May need to remap/restripe data.
- Backup is one snapshot. HSM is many versions?

Improved support for multi-clustered environments (including QoS)

- Discussion Comments
- Multi-Clustered Environments (w/ QoS)
    - Rolling upgrade requirements – are 2 Lustre versions simultaneously required?
    - It's a problem to have to start with a 1.4 client to upgrade to 1.6 to get a client that can talk to a 1.4 server. A clean install of a 1.6 client can't talk to 1.4 But there are common use cases that require this.
    - More than 3 versions of compatibility are needed.
    - Resources spent on compatibility are not working on new features.
    - Any 1.8 should work with any 1.6.
    - ZFS-CMD upgrade, LNET IPv6 are big looming change. IPv6 presents a wire format change. This will require major work if we must interoperate between IPv4 and IPv6.
    - Lustre will negotiate features between client and server
    - Are routers dedicated? Not always
- QoS
    - NRS needs to be coordinated across servers.
- Future HPC systems
    - Liblustre could be ported to new OSs
    - TCP can run over most interconnects. This is a fallback strategy
    - Can Lustre-compatibility be specified in the procurements?
    - Can't necessarily limit competition by specifying Lustre, especially now that Lustre is owned by Sun. Could get into competitive conflicts.

Data Integrity

- Discussion Comments
  - lfsck improvements
  - May need to modify standards to get Lustre features supported
  - Customers need to know the cost to keep same level of performance
  - Lustre team should be able to do testing fairly soon
  - 25% performance hit would likely be too much. Might have to look at other options
  - Lustre team will be able to improve ZFS, look at ext3 track record
  - All code needs to be parallel across cores, since core speeds are not increasing. Checksums are a good example of this.
  - Integrity checking must be done in Lustre, since we support so many types of hw. Can't just rely on exploiting one integrity feature of one vendor's hw
  - Lustre RAID – performance and recovery issues. There are substitute technologies. Write coherence to RAIDed servers is the problem.
  - With failover issues, and we don't know what disk environment looks like, LRAID is a nice option. Addressing concerns about failover would help.
  - Have started architecting LRAID. Came up against issue of locking the stripes sequentially to avoid cascading abort problem. This produces a performance problem. Either abort problem or performance problem. May need to put this on Lustre-devel.
  - Maybe hold a workshop on this.

`

Evolve Lustre towards a more community driven development model

- Discussion Comments
  - Lustre should leverage larger community
  - Trying to get Livermore's patches back into code tree.
  - Need joint copyright assignment in order to do this
  - The controller of the code has to be able to redistribute it. That's why joint copyrights are needed.
  - Do not want to be forced into a change like ZFS without some community discussion.

Support for Very/Ultra-Large Large Clusters and WAN

- Discussion Comments
  - Need to use proxy servers to
  - Must be thinking about scaling to 100K+ clients
  - Need a Collectl project at ORNL
  - Recovery protocol between MDS and 1000's of OSSs will not scale
  - The pinger may have some scalabilty issues.

HSM

Lustre HSM Requirements (1/2)

- **An HSM extension for Lustre**
  - To inter operate with existing storage systems
  - No strong binding with external storage
    - ☛Basic copy-in, copy-out must work with a simple user space tool
- **Provide basic features**
  - Cache miss, archive, purge, transparency
  - Can be used as backup

Lustre HSM Requirements (2/2)

- **All files are always visible in the file system, but a file can reside:**
  - On primary storage (Lustre)
  - On the backend storage
  - On both
- **Metadata (size, …) are always up-to-date**
  - Add a migration status flag
- **Scalable and parallel**
  - Lustre HSM must have a small impact on Lustre performances
  - Target is to impact Lustre performances only when data are not in Lustre (time to bring back data when a cache miss occurs)

Lustre HSM Requirements (2/2)

- **All files are always visible in the file system, but a file can reside:**
  - On primary storage (Lustre)
  - On the backend storage
  - On both
- **Metadata (size, …) are always up-to-date**
  - Add a migration status flag
- **Scalable and parallel**
  - Lustre HSM must have a small impact on Lustre performances
  - Target is to impact Lustre performances only when data are not in Lustre (time to bring back data when a cache miss occurs)

Inside Lustre HSM (2/2)

- **Use of pre-migration**
  - Automatic
  - On demand: with a user space command
- **File system space management is either:**
  - Automatic
    - At OST level
    - At FS level (MDT)
  - On demand: Based on a provided list of files
- **Purge method**
  - Keep start/end of FID on disk
  - At OST level (objects)
  - At FS level (all file)

Lustre HSM Components (1/3)

- **Initiators**
  - A node placing a migration request with a coordinating node
  - Handle cache misses
- **Coordinators**
  - A service coordinating migration of data
  - Activate agents to move data
  - Manage multiple requests (2 types of requests: implicit and explicit)
  - Send callbacks to initiators
  - Support migration cancel requests

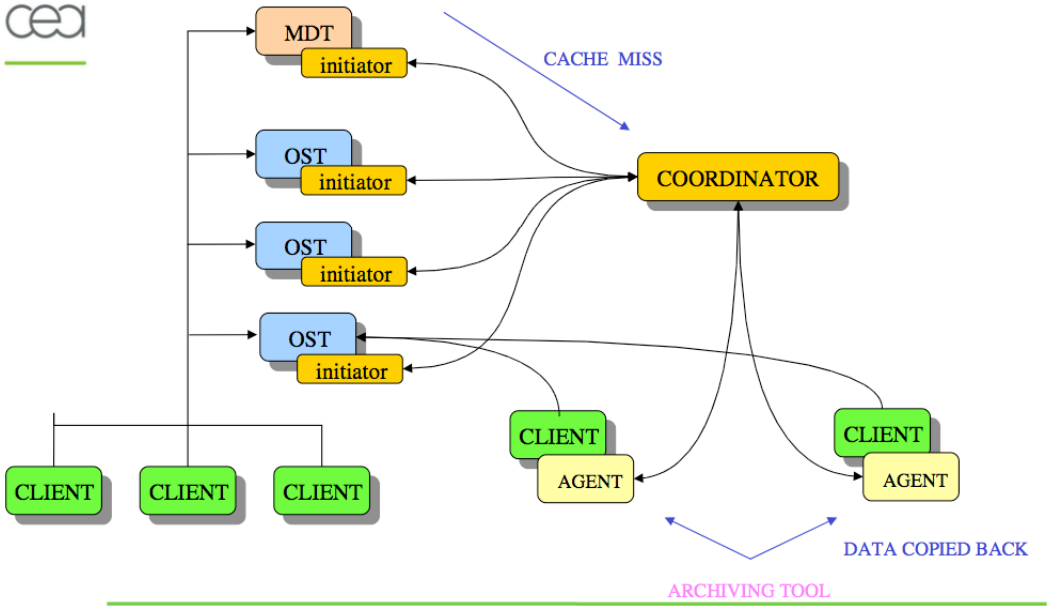
Lustre HSM Components (2/3)

- **Agents**
  - A service used by coordinators to move data, cancel such movement and remove external storage files
  - They invoke HSM tool
- **HSM Tool**
  - An external storage is defined by a label and associated to a copy tool
  - A user space tool used to interface to the external storage
  - Copy-in, Copy-out, Remove, Cancel (optional)
  - Multiple files can be managed by one request, tool can choose to regroup them in one archive

Lustre HSM Components (3/3)

- **Space Manager**
  - A service in charge of pre-migration and space management
  - Use of migration policies
- **Scanners**
  - A tool used to generate list of files without going through the namespace
  - Depend of the MDT backend

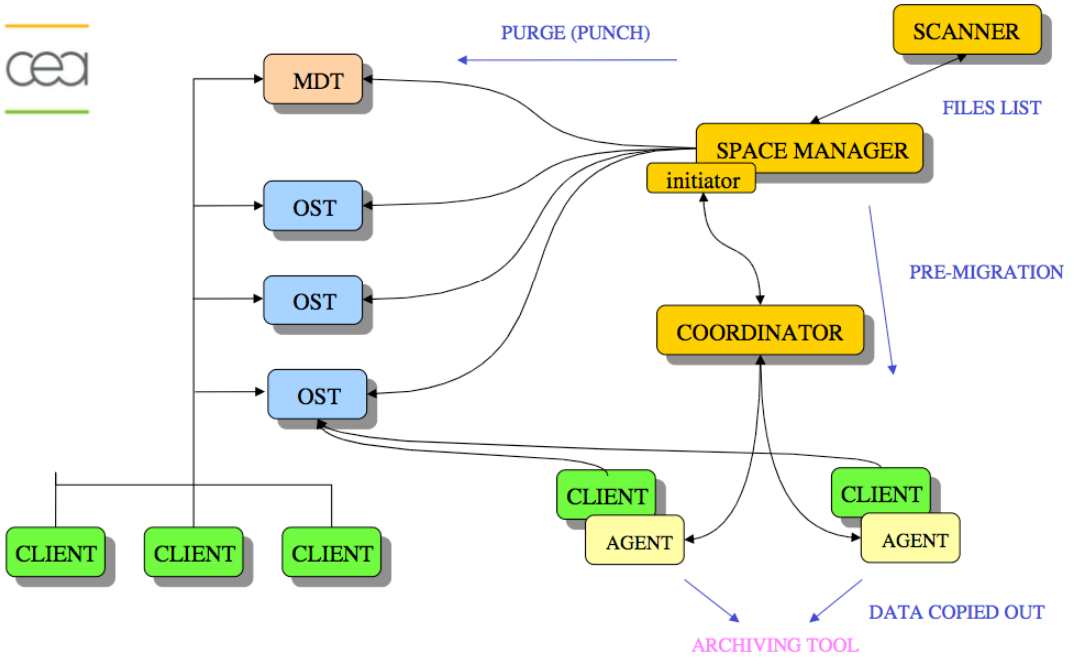
Migration Architecture

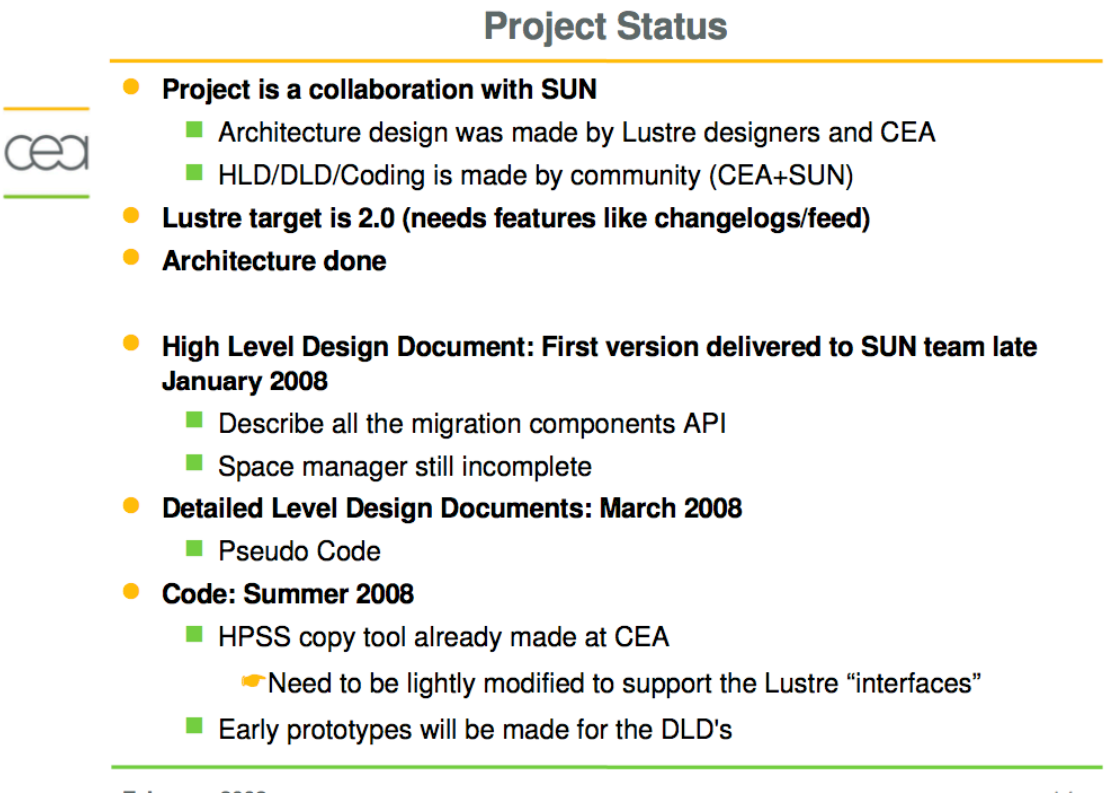


External HSM requirement

- **A userspace command able to**
  - Copy from posix (Lustre) to HSM
  - Copy from HSM to posix (Lustre)
  - Remove a file in HSM
  - Cancel a transfer (optionnal)
  - No Lustre knowledge is needed in the HSM
    - ☛Lustre access is made through a hidden path (/mnt/.lustre/FID/...)
  - Manage a data transfer cursor
- **HSM namespace based on Lustre FID**
- **A reference to HSM object ID and a version number (returned by HSM) is kept in Lustre**
- **Support of Named Attributes in HSM will allow**
  - Backup of some file attributes in HSM (at migration time)

Space Management Architecture





- Discussion Comments
  - CEA HSM solution will store Lustre files by FID in HSM system. Lustre will need to store HSM FIDs with each file.
  - Braam pointed out that extended attributes to store HSM FiDs, striping info, etc need to be common across file systems.
  - How to restore from HSM if it doesn't have a copy of the file system name space?
  - The scanner tool should be changed to use the LLOG for LRU file info. Could have HSM policies based on size or age. LRU is not sufficient.
  - NRL is getting changelog that could be used as a fast query tool.
  - Another policy dimension. Users to create their own policy? A new mechanism is needed to implement user policies. Need database, not config files.
  - Don't have backup/restore use cases. Need to define these.
  - ORNL needs to review HLD. They may want to contribute to some of the development
  - JC: space manager, policy database – could use help here.
  - Anyone can add use cases to arch wiki.
  - JC to summarize use cases.

Meeting Thoughts – Eric Barton

Stability

- Code Ownership & Coverage
  - No dark corners
  - Clear, documented internal APIs
  - Subsystem Experts
- Process
  - Branch management
  - Concurrent feature development
- Fulfill Expectations
  - Build on solid ground
  - Believable roadmap
  - Interoperability
    - Enumerate required use cases
    - Limit complexity
- QE
  - Regression test automation
  - Customer site

Meeting Close – Peter Bojanic

Participants

- Focus was **big**, **strategic**, and **HPC**
  - > Invitations based on merit
  - > Choices made largely by ORNL and Sun
- Future participants
  - > Suggestions of some commercial representation
  - > How do we determine who to invite?

What We Accomplished

- Lots of fruitful, constructive discussion
  - > Excellent level and quality of engagement
  - > Common understand of each organizations' vision and priorities for Lustre
  - > Community ranking of priorities
  - > Sufficiently deep dives into hottest topics
- Sun engineers and managers more closely engaged
- Community leadership emerging

What We Could Have Done Better

- Somewhat more near sighted than anticipated
  - > Not entirely surprising
  - > Not necessarily a bad thing
- Community responsibility
  - > Only eeb is going home with any serious homework
  - > No substantially greater community ownership

Community Development

- Gel an LCE Council
  - > This group represents the founding members
  - > Who are the leaders that will step forward?
- Define a mandate for the group
  - > Responsibilities
  - > Authority
  - > Accountability
- Define criteria for membership

Community Forums

- Mailing list
  - > lce-council@lists.lustre.org
  - > Public subscriptions with approved posters
- Lustre wiki
  - > http://wiki.lustre.org
  - > Make process documentation public for review
- Architecture wiki
  - > http://arch.lustre.org
  - > Elaborate requirements on the web site
  - > Use wiki for discussion, change tracking, notification

Requirements Management

- Publish of community priorities
  - > We'll be asking for your permission
- Tied to the Lustre road map
  - > In an explicit way
  - > That is traceable
- Acceptance criteria
  - > Build on established standards
  - > Ask community for input

HSM

- Most ambitious Lustre community project ever undertaken
- HLD is published
  - > On lustre-devel@lists.lustre.org
  - > Feedback is highly encouraged
- CEA may consider filing a bug
  - > Track input via comments
  - > Version HLD via patch management
- What are next steps?

Next Steps

- Feedback from Lustre Group at Sun
  - > This was entirely worthwhile -- let's do it again
- Next meeting
  - > Six months or one year?
  - > Is Sun Burlington campus generally acceptable?
- Follow up discussions
  - > What would you like to see?

Credits

- ORNL – Al Geist, Shane Canon
- Peter Braam
- Dan Ferber
- Lindsey Stack
- Bryon Neitzel
- Eric Barton